TARTU UNIVERSITY

FACULTY OF SOCIAL SCIENCES

NARVA COLLEGE

STUDY PROGRAM “INFORMATION TECHNOLOGY SYSTEMS DEVELOPMENT“

Liubov Ustinova

COLLECTING AND PROCESSING UNSTRUCTURED TEXT DATA ON ENDANGERED BIRD SPECIES IN ESTONIA

Bachelor’s thesis

Supervisor: Nicolai Morozov

Narva 2025

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Liubov Ustinova,

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose: “Collection And Processing Unstructured Data On Endangered Bird Species In Estonia”, mille juhendaja on Nicolai Morozov, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Liubov Ustinova

17.05.2025

TABLE OF CONTENTS

TABLE OF CONTENTS 3
LIST OF TERMS AND ABBREVIATIONS 4
INTRODUCTION 5
RESEARCH TASKS AND GOALS 5
1. BACKGROUND 7
1.1. Existing Solutions And Their Limitations 7
1.2. Large Language Models And Their Potential In Raw Text Data Processing 11
1.3. Importance Of The Study 11
2. DATA AND METHODS 13
2.1. Data Sources 13
2.2. Method Limitations And Challenges 13
2.3. Workflow 14
2.3.1. Data Collection 17
2.3.2. Unstructured Data Cleaning 18
2.3.3. LLM Usage 18
3. RESULTS 19
3.1. Criteria And Methods For Technology Reliability Evaluation 19
3.2. Test Results 19
3.2.1. Assessment of the quality of the algorithm's performance 20
3.2.2. Assessment of the quality of the output data 21
CONCLUSION 30
RESUME 31
REFERENCES 32

LIST OF TERMS AND ABBREVIATIONS

Data completeness analysis - The process of evaluating a dataset to determine whether all necessary data points are present and accounted for. This analysis identifies missing or incomplete data entries.

EELIS - Estonian database used for centralized managing environmental or ecological data.

Evaluator - Person that assesses or measures the quality, performance, or value of an entity, such as a piece of work, a product, or a system.

gpt-4o-mini - Smaller, specialized version of GPT-4 designed for optimized performance or specific tasks.

LLM - Type of AI model called a "Large Language Model" that processes and generates human-like text.

NLP - Natural Language Processing, a field of AI focused on the interaction between computers and human language, including understanding, interpreting, and generating text.

OCR - Optical Character Recognition, a technology that converts different types of documents, like scanned paper documents or images, into editable and searchable data by recognizing text characters.

Ornithology - The scientific study of birds, including their behavior, ecology, and conservation.

pandas - Python library for data manipulation and analysis, offering data structures like DataFrames.

Python - High-level programming language known for its readability and versatility in various applications.

Selenium - Python library for automating web browsers, commonly used for testing web applications and data parsing.

Species categories of protection - Classifications used to indicate the conservation status and protection level of different species, such as endangered or vulnerable.

Taxonomy - The scientific discipline of classifying organisms into a structured hierarchy based on shared characteristics and evolutionary relationships.

INTRODUCTION

Technological advancements, including large language models (LLMs), are transforming various fields. In ornithology, LLMs can streamline data collection, saving time for professionals. This thesis project proposes exploring the use of LLMs for one of the field's fundamental tasks: data collection for marketing, educational materials, and scientific presentations.

Ornithology, the study of birds, plays a vital role in biodiversity research and conservation. Birds serve as key environmental indicators, their decline signalling broader ecosystem distress. In Estonia, **114 bird species** fall under protected categories I, II, and III in EELIS [1]. Category I covers critically endangered species facing severe threats, Category II includes species at risk without intervention, and Category III consists of currently stable species that may decline under persistent threats. Understanding these categories and their important place in Estonian ecology leads to realising how important access to complete and accessible data on them is in such conditions.

This is why it is essential to ensure that professionals and enthusiasts can access comprehensive and reliable bird species data. Currently, resources in Estonia, such as EELIS, KESE, PlutoF, and eElurikkus, provide similar data. Yet, each has limitations in providing quick access to general information like species descriptions and major threats. eElurikkus relies on PlutoF's data, while PlutoF itself, as a biological data management platform, poses challenges for ornithologists not directly involved in research to navigate the Taxonomy section. Furthermore, species descriptions in some cases are limited to one or two sentences which may be insufficient for certain tasks [2]. The EELIS database, used extensively by ornithologists and supporting KESE, is the most convenient; however, data on threats or descriptions of some species are also lacking or incomplete. This thesis project aims not to replace existing tools but to offer a solution that can address existing or future data gaps by leveraging conservation strategy studies and ornithological articles, ultimately preserving the invaluable resource of time for everyone involved.

RESEARCH TASKS AND GOALS

Relying on the problem identified in the previous section, the following main objectives of this study can be highlighted:

- Verification of the hypothesis concerning the relevance and novelty of the project

  The importance of precise and thorough bird data in conservation plans makes it essential to confirm the project's relevance in addressing data gaps that are not sufficiently covered by Estonia's current resources. This goal guarantees that the project's emphasis on processing ornithological data using LLMs is up to date, innovative, and makes a significant contribution to the field.

- Creation of the workflow

  Effective use of the capabilities of large language models in ornithology requires a clearly defined workflow. This goal focuses on developing a methodical strategy for integrating LLM usage and setting up the data collection process. This guarantees that ornithological data can be collected and processed using the technology in an efficient manner.

- Justification of the approach's correctness

  Since the project suggests an innovative use of LLMs in ornithology, it is essential to show that the selected method is accurate and dependable. This entails confirming that the technologies can process unstructured data and offer precise and practical insights, meeting the requirements of conservation initiatives and reducing the difficulties presented by the data management systems in place. The effectiveness and dependability of the technology are assessed by the analyzing of the results of evaluation of output data by professional ornithologists. Their evaluation will include grading the output species descriptions and dangers with grading scale 0-10 (where 0 is completely inaccurate, and 10 is information presented is complete and up-to-date).

- Analysis of the obtained results

  After implementing the workflow and processing data with LLMs, a thorough outcomes analysis is required. This objective assesses the effectiveness and quality of the data processed by LLMs, evaluating whether it meets the necessary standards for ornithological studies and conservation strategies, and compares it to data prepared by human experts.

- Presentation and formatting of the obtained results

  Lastly, to make sure ornithologists and enthusiasts can easily use the data processed by LLMs, it must be presented effectively. In order to accomplish this goal, the data must be arranged, formatted, and presented in a way that makes sense for tasks involving data analysis and visualization.

Thus, the ultimate goal of the study is to demonstrate the use of LLM for processing unstructured data and to observe the quality of the output text descriptions of species and their threats.

1. BACKGROUND

This thesis's background section establishes the basis for exploring how large language models (LLMs) can be integrated into ornithological data processing, emphasizing both the advantages and disadvantages of the data management systems in use today. It examines current approaches and their drawbacks, with an emphasis on Estonian resources that, although useful, have accessibility issues and insufficient threats and species descriptions. By effectively processing vast amounts of ornithological literature, large language models (LLMs) offer promising ways to fill in current data gaps. This is demonstrated in the discussion of LLMs and their potential in raw text data processing.

Lastly, the section clarifies the need for advanced techniques of data collection and utilization in ornithology, which will ultimately support more successful conservation strategies and make crucial information easily accessible to both professionals and enthusiasts. This will greatly aid efforts to preserve biodiversity.

1.1. Existing Solutions And Their Limitations

In order to evaluate the necessity of establishing this project, it was important to study existing analogues, assess their advantages, and identify their vulnerabilities. To this end, an ornithological survey was developed, comprising a series of questions and corresponding answer choices.

For the purpose of the study, contact information for ornithologists, members of the Estonian Ornithological Society (Eesti Ornitoloogiaühing), Kotkaklubi, as well as Tartu University zoologists, was meticulously gathered. This involved a manual review of existing ornithological reports, followed by an extensive internet search to locate their email addresses, as both the Ornitoloogiaühing and Kotkaklubi provide public access to the contact details of only a limited number of their members. This step resulted in accessing 21 contacts (while there are about forty active ornithologists or directly related to scientific ornithology in Estonia (based on data on Kotkaklubi and Eesti Ornitoloogiaühing participants and the number of contacts collected). Below, the description of the survey form questions and received results can be viewed.

In the context of this research project, a survey was conducted among professional ornithologists and zoologists to assess the existence and necessity of an instrument for dealing with the lack of data in the currently existing databases for endangered bird species in Estonia. The survey aimed to highlight the potential benefits of establishing such a resource. A total of 7 responses were collected. Below is a detailed description of the survey questions and the responses received.

Survey Questions and Responses:

1. Have you encountered a lack of a unified source of data on endangered bird species in Estonia in your work?

Most respondents answered "Yes," indicating that the absence of a centralized data source poses a challenge in their work. Notably, those who answered "No" stated that they currently use existing resources like EELIS, KESE, and PlutoF, which partially fulfill their needs.

2. If your answer to the first question was "No," do you currently have access to similar information or databases? (If yes, please specify the source).

Those who answered "No" to the first question specified that they use sources mentioned above, demonstrating a reliance on multiple existing databases.

3. If your answer to the first question was "Yes," do you feel that having a resource obtaining full data would significantly save your time?

Respondents who initially identified the need for a unified source expressed that such a resource would indeed significantly reduce the time spent searching for data.

4. If you had access to such data, how would you use it in your activities? (Select all applicable options):

- In-depth species research

- Development of conservation measures

- Preparation of scientific articles and presentations

- Creation of educational materials

The majority indicated that they would utilize the data to enhance species research, develop effective conservation strategies, and prepare scientific publications.

Basing on their responses, the list of the main ornithology data systems was created: EELIS, KESE, and PlutoF, intended to support the work of ornithologists. However, each of these resources is hindered by noticeable limitations, particularly in terms of access to comprehensive species descriptions and details on major threats. Below, these limitations are analyzed and described to provide a clearer understanding of their impact on ornithological research and conservation efforts.

1. EELIS database

For information on Estonian bird species, ornithologists rely heavily on the EELIS database, which is well known for its ease of use. Because of its extensive data coverage, this platform acts as the foundation for other databases, including KESE. The primary shortcoming of EELIS, however, is its incapacity to offer comprehensive information on the threat levels and descriptions of certain species, which may be either absent or insufficient (relying both on ornithologists feedback in completed survey and data completeness analysis conducted, results of which are available on the pie chart below). Since users are unable to access complete information required to assess the unique threats that each species faces, this deficiency directly impedes thorough species assessments and the creation of conservation strategies.

Below the pie diagram (Figure 2) empathizes that **although EELIS maintains 57% of its data with adequate detail (66 species), a concerning 42% (48 species) related to endangered birds, classified within categories I, II, and III (encompassing 114 birds earmarked for protection), lacks sufficient descriptions**. This incomplete depiction of threatened species restricts ornithologists and conservationists who depend on thorough and accurate data for decision-making.

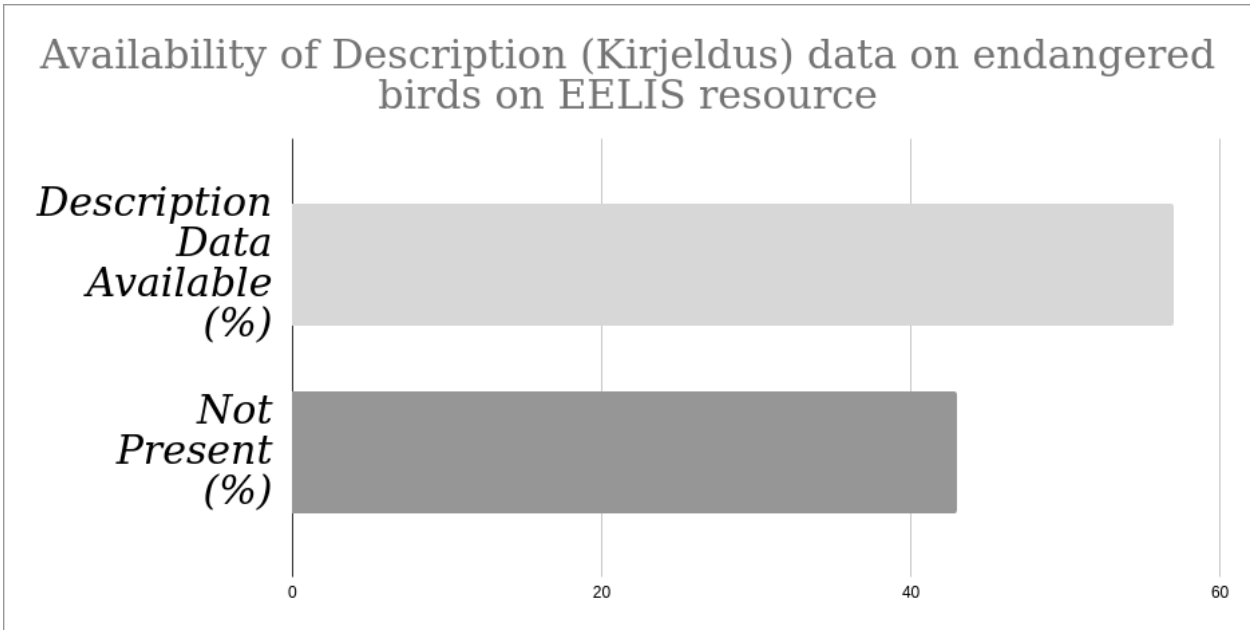

Figure 2. *Chart displaying the completeness of threat and description data in the EELIS database for endangered birds*

2. KESE

The KESE database, which is a derivative of EELIS, has similar challenges. It shares its limitations because it heavily depends on data that was passed down from EELIS. Because of its dependence, any inaccuracy or lack of data in EELIS is reflected in KESE. Because of this combined effect, there are significant information gaps that prevent ornithologists from fully understanding the ecological status and conservation requirements of different species.

3. PlutoF

Known for its ability to manage large amounts of ecological and taxonomic data, the PlutoF database is a comprehensive platform for managing biological data. For ornithologists, especially those who are not directly involved in biological research, it presents serious difficulties. A key component for obtaining important species information, the platform's Taxonomy section is frequently challenging to use. For users who need simple data access but might not have the specialized knowledge required to function effectively in such a complex system, this complexity poses a significant obstacle.

Navigating PlutoF's Taxonomy section can be a daunting task for many ornithologists, particularly those who are more focused on fieldwork than computational tasks. Its complex structure requires a strong grasp of taxonomic data management, which results in a significant time commitment and a steep learning curve. This makes it difficult to retrieve information quickly, which makes it more difficult to collect important data, especially when there are pressing conservation challenges.

Furthermore, the short species descriptions in PlutoF also present a big problem. These entries, which are frequently limited to one or two sentences, are inadequate for tasks that call for in-depth biological knowledge, like species profiling, ecological impact assessments, and the creation of educational materials. This lack of depth impedes users from achieving a comprehensive understanding of species-specific traits and ecological roles, which are essential for informed research and strategic conservation planning.

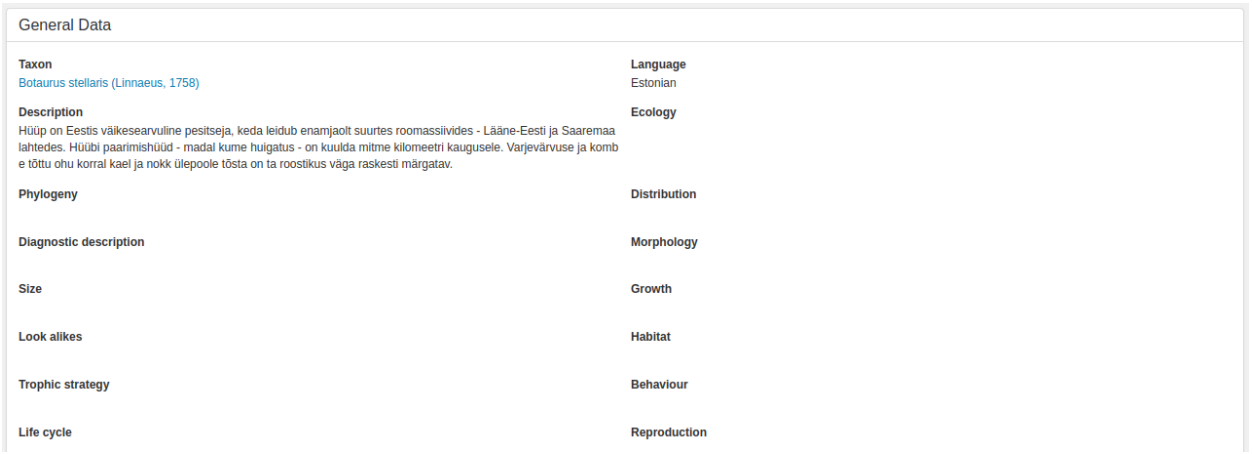

Figure 3. *Sample of limitations of the endangered birds descriptions at PlutoF resource for hüüp (Great Bittern)*

In conclusion, the examination of the three ornithological databases currently in use in Estonia - EELIS, KESE, and PlutoF - shows several limitations, especially with regard to providing thorough, centralized, and easily navigable access to vital information on endangered bird species. Although these resources provide useful information, their disadvantages, such as missing data, reliance on overlapping datasets, and difficult navigation - highlight the ongoing difficulty ornithologists have in obtaining complete and actionable data.

According to the completed survey, 6 out of 7 of participants acknowledge the issue of incomplete or fragmented data, highlighting the need for a more reliable solution. By providing a solution that can aid in creating or updating an integrated and easily accessible data resource, the proposed project seeks to address these limitations. It may be able to close highlighted knowledge gaps, which would help with better species research, conservation strategy development, and the production of educational resources related to Estonia's endangered bird species protection.

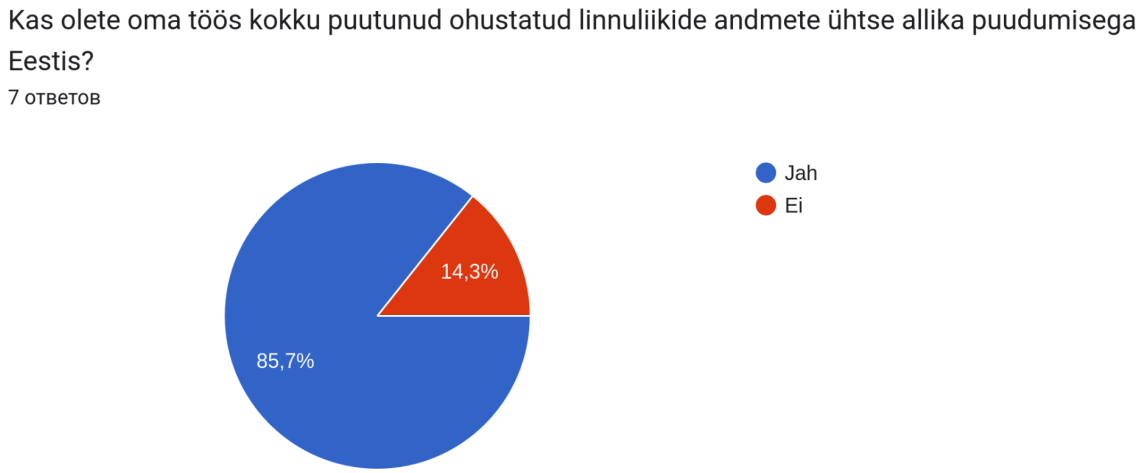

Figure 4. *Pie chart displaying responses to the question on data deficiency for endangered birds in Estonia*

1.2. Large Language Models And Their Potential In Raw Text Data Processing

Recent large scale language models (LLMs) improvement has introduced transformative tools capable of revolutionizing various tasks, including text summarization, which is especially important for managing and utilizing unstructured data. In the context of the described project, LLMs present an opportunity to enhance speed and quality of data collection and processing, allowing professionals to focus on complex and specialized tasks by leveraging their automated summarization capabilities. Current ornithological resources may face challenges related to data accessibility and completeness. Current project, accessing the opportunities LLMs offer a solution to these issues by effectively summarizing extensive ornithological literature and resources, filling the gaps where traditional databases may fall short.

Text summarization, on which the major part of the project relies, is a part of Natural Language Processing (NLP) that can distill large volumes of detailed scientific text into concise, informative summaries. Recent studies have demonstrated the effectiveness of LLMs, like the MPT-7b-instruct and OpenAI ChatGPT models, in performing both abstractive and extractive summarization tasks. These models not only capture the essence of complex texts but also maintain the key informational components necessary for diverse applications [6].

Text summarization through the LLM usage is performed directly through prompt-based querying of LLMs, allowing for dynamic interaction with the data and customization of extraction criteria. As noticed in the work of Schilling-Wilhelmi et al. [7], the task-specific adaptation of LLMs can be effectively achieved through prompt engineering, where simple zero-shot or few-shot strategies are often sufficient to extract structured data from unstructured sources.

The ability of LLMs to generate succinct and coherent summaries based on ornithological data is crucial for enhancing information retrieval and content generation in the field. By employing powerful algorithms for text summarization, LLMs support ornithologists in accessing precise data efficiently, thereby advancing conservation efforts and ensuring the effective dissemination of scientific knowledge.

1.3. Importance Of The Study

The importance of this study is underlined by the gaps identified in the completeness of data identified through the results of surveys and extensive analysis of existing Estonian ornithological databases, such as EELIS, KESE, PlutoF poses challenges The survey found that most experts face problems due to fragmented information resources. This prevents access to detailed, comprehensive and reliable information on endangered bird species in Estonia.

However, this study doesn't just identify the problem. It also offers effective solutions to address these issues. The proposed solution has the potential to efficiently fill these data gaps by using large-scale language models (LLM) for automatic text summarization and data processing. **Within the project, key LLM capabilities were applied to automatically extract relevant text segments (e.g. sections on habitat, threats, population changes) from raw documents and then generate detailed and structured textual descriptions of species and threats. It was the potential of LLM to automate the analysis of textual data and fill information gaps that was the determining factor in choosing this method, as existing systems do not provide the necessary completeness of data and manual processing is not scalable or efficient for large amounts of information.**

Moreover, the proposed solution not only addresses data integrity issues, but it also **saves a lot of time for ornithologists and experts**. By **automating the process of collecting and summarizing data, businesses can redirect their focus to higher-level analytics and strategic planning.** This will help increase the efficiency and effectiveness of conservation efforts.

Ultimately, this study offers an approach to improve data accessibility and utility in ornithology, supporting more informed and timely conservation strategies. By improving the quality and availability of essential information, it contributes significantly to biodiversity preservation efforts, aligning with broader ecological goals and enhancing the capacity for wildlife protection in Estonia and beyond. Through the integration of advanced data processing methodologies, this project takes a crucial step toward more sustainable and resilient conservation outcomes.

2. DATA AND METHODS

2.1. Data Sources

This project's data collection and management on endangered bird species rely on various comprehensive sources. Each source is vital for constructing a detailed and actionable dataset to support conservation efforts.

The Estonian Governmental Resources ("Riigi Teataja") is the first key data source. This source provides comprehensive lists of vertebrate species categorized under various conservation threat levels (I, II, and III) but is not initially segregated by species type. The data forms the foundational dataset for the project. By broadly targeting vertebrate species, the project later refines this data to focus specifically on birds (Linnud), which is central to the study. URLs from this site are parsed to download raw data, facilitating a broad overview of species needing conservation action in Estonia.

Another important source is the Estonian Environmental Information System (EELIS), which offers individual species pages with granular details about endangered bird species. This enrichment process enhances the dataset with targeted information, filling data gaps about species-specific conservation needs, current threats and whether any conservation strategies are available. Selenium is used to automate the retrieval of this detailed data.

Online resources accessed via Google Search constitute a data source, focusing on PDF documents containing potential conservation action plans or related environmental documentation. Accessing these documents allows the augmentation of the dataset, specifically focusing on birds without existing preservation strategies. This integration of external documents provides additional resources to fill descriptive gaps and identify threats.

Each data source collectively supports the project's aim to synthesize a comprehensive dataset to inform conservation strategies for endangered bird species, particularly within Estonia.

2.2. Method Limitations And Challenges

There are several key issues and limitations that arise during the process of data collection, management, and analysis, specifically concerning the study of endangered bird species in Estonia.

One important limitation of this project, which directly affects the completeness and quality of the results obtained, is its dependence on the availability of a sufficient amount of relevant baseline data for each specific bird species. As the developed workflow is based on searching, retrieving and processing existing unstructured text documents (such as conservation plans, reports and articles found via Google Search or referenced in EELIS), the ability of the system to provide detailed descriptions and analyses of threats is directly proportional to the availability and information content of these sources. In cases where such documents are unavailable or contain minimal information for a particular bird species, the system is unable to generate comprehensive summarisations. This results in gaps in the final dataset, which manifests as incomplete records or “NA” (Not Available) designations for missing information. Thus,

although the project is effectively processing available textual data to fill identified information gaps, its success for a particular species is limited by the existing amount of unstructured information published about it.

Another challenge relates to the linguistic aspect of the project. The fact that all primary literature and resources are in Estonian presents a significant obstacle, even if the language is supported by models like GPT-4o, despite its capabilities the model may provide less accurate results compared to processing documents in more widely used languages due to potential nuances and specific terminology found in Estonian ornithological texts. This can affect the accuracy of summarization and extraction tasks, potentially leading to the inclusion of irrelevant information or the omission of critical details necessary for comprehensive analysis.

Evaluating the correctness of the summaries and ensuring that they only include relevant information without any extraneous data adds an additional layer of complexity. The limitations in natural language processing models, such as possible misinterpretations of context or subtleties in the language, can challenge the effective summarization of documents. This necessitates rigorous validation mechanisms and manual oversight to verify that the extracted summaries are both accurate and pertinent to the conservation objectives.

Finally, the process requires intensive data cleaning and preparation. Dealing with unstructured and inconsistently formatted data requires a systematic approach to ensure reliability and utility. Extracting meaningful insights demands considerable effort in filtering, cleaning, and aligning data, which can be resource-intensive. Each step, from the initial data scraping to the detailed extraction using large language models, presents potential points of failure that require careful monitoring to ensure the usability and corectness of the collected data.

2.3. Workflow

This section presents an overview of the systematic approach undertaken to process and analyze data related to endangered bird species in Estonia. The workflow is designed to transform raw data into structured insights, facilitating a comprehensive understanding of bird conservation needs. This multi-step process is depicted in the sequence diagram below, which illustrates the interactions and dependencies among various project stages and in the text of section chapters, which describes all process steps in detail from initial collection to final analysis.

The main contribution of this project lies in the creation and implementation of an automated workflow that effectively combines the capabilities of various existing tools and data sources to address the task of collecting, processing and summarising unstructured ornithological texts. While components such as large language models (LLMs), specialised libraries and tools for automating web access (Selenium), text manipulation (Requests, BeautifulSoup, fitz, pdftotext tool, regular expressions) and API access (OpenAI API) are off-the-shelf technological solutions, the described work consisted of designing and writing a control scripts in Python that orchestrate their interaction. This solution implements the sequence of steps described in this section, providing automated data collection, data cleaning, extraction of relevant fragments and, most importantly, the use of LLM (gpt-4o-mini) to analyse and create structured summaries based on the information found.

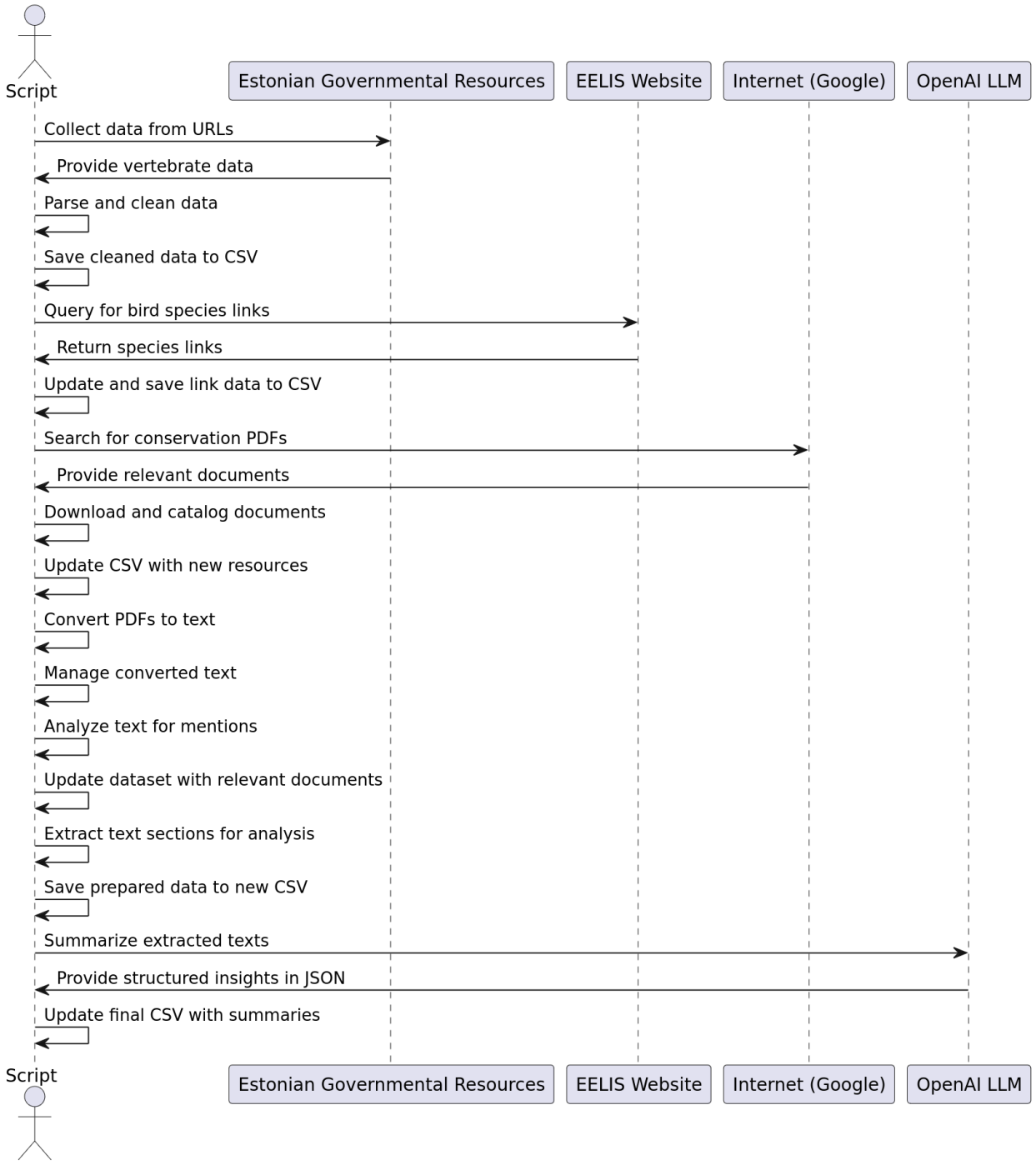

Figure 5. *Sequence diagram illustrating the systematic workflow for processing and analyzing data on endangered bird species in Estonia*

Process Interactions and Workflow Description

The **Script** automates a sequence of data collection, processing, and analysis tasks, interacting with various external and internal components to achieve structured data insights.

- **Data Collection from Governmental Sources**
    - The Script initiates the process by querying the **Estonian Governmental Resources (Govt)** for vertebrate data.
    - The Govt system responds by providing relevant datasets.

    - The Script then parses and cleans the received data, ensuring consistency and removing unnecessary information.
    - Finally, the cleaned data is saved into a CSV file for further processing.
- **Retrieving Bird Species Links from EELIS**
    - The Script interacts with the **EELIS Website** to retrieve bird species links.
    - EELIS responds by returning a set of relevant species links.
    - The Script updates and saves the link data into a CSV file for further reference.
- **Searching for Conservation Documents on Google**
    - The Script queries **Google** for conservation-related PDFs.
    - Google responds by providing access to relevant documents.
    - The Script downloads and catalogs these documents, storing them systematically.
    - It then updates an existing CSV file with references to newly acquired resources.
- **Text Extraction and Preprocessing**
    - The Script converts downloaded **PDFs to text** for easier analysis.
    - It manages and organizes the converted text data.
    - The Script analyzes the text for **mentions of relevant topics**, identifying key conservation references.
    - The dataset is updated with these identified documents.
- **Text Segmentation and Data Preparation**
    - The Script extracts meaningful **text sections** for analysis.
    - These extracted segments are saved in a new CSV file for structured data storage.
- **Summarization with OpenAI LLM**
    - The Script sends the extracted text to the **OpenAI LLM** for summarization.
    - The LLM processes the input and returns structured insights in **JSON format**.
    - The Script takes the generated summaries and updates the final **CSV file** with concise, structured insights.

Overall Interaction Summary

- The **Script** acts as the central processor, automating interactions with external data sources (**Govt, EELIS, Google**) and utilizing **OpenAI LLM** for summarization.
- Internally, the Script manages **data parsing, cleaning, saving, analyzing, extracting, and updating datasets** to ensure structured and refined information is obtained.

- The **CSV files** act as intermediate storage at multiple stages, allowing the system to track and store processed data systematically.

This workflow automates data collection, processing, and insight generation, ensuring a structured and well-documented approach to conservation data management.

2.3.1. Data Collection

The data collection process for this project is structured in a series of strategic steps, each building upon the previous, to ensure a comprehensive and accurate dataset.

The initial step involves automating the collection of vertebrate species data from the Estonian Governmental Resources ("Riigi Teataja"). This step acts as the foundation for the entire project by gathering comprehensive lists of species categorized under various conservation levels. The data includes all vertebrates and is later refined to focus specifically on birds. This automated process utilizes Python scripts and essential libraries such as Requests and BeautifulSoup to download and parse HTML content, storing the cleaned results in a structured CSV file.

The next step is to enhance the data collected about bird species using the Estonian Environmental Information System (EELIS), which builds on the original dataset. In order to automatically query EELIS with species names from the original dataset and retrieve comprehensive links to species-specific pages, this phase makes use of Selenium WebDriver. In addition to offering a more focused perspective on birds, which is essential for the project's scope, these links enhance the dataset by enabling an analysis of current data gaps.

Following the identification of the bird species lacking documented strategies, Google Search is used to perform focused web searches in order to obtain more documents. This entails looking for PDF files that include relevant documentation or possible conservation action plans. Utilizing Python in conjunction with libraries such as Requests and Googlesearch-Python makes it easier to obtain these outside sources, adding more relevant information to the dataset. In order to fill in the data gaps identified in earlier phases, this step is essential.

Next, the project automates the conversion of collected PDF files into text format to streamline subsequent analysis. Utilizing the pdftotext tool through Python's subprocess module, PDFs are converted to text files, making them more accessible for analysis using natural language processing techniques. This step prepares the dataset for detailed content analysis, building on the wealth of information gathered in earlier steps and enhancing it for further examination.

In the following step, the project determines the most relevant document for each bird species by analyzing text extracted from PDFs for frequency of species mentions. This involves using Python libraries such as fitz (PyMuPDF) and regular expressions to identify which document best represents each bird species based on name occurrence. This analysis sharpens the dataset, extracting the most pertinent information and ensuring that future analyses are focused on the most informative resources identified in previous stages.

Finally, the project uses a large language model (gpt-4o-mini) to extract and summarize relevant text sections from the gathered documents. This step synthesizes complex information into structured summaries focusing on bird habitats, threats, and population changes, leveraging the OpenAI API for sophisticated analysis and content cleaning. This final step ties in all previous data collection steps, using the rich dataset constructed through earlier steps to deliver insights necessary for effective conservation planning. These detailed summaries culminate the data collection process, transforming raw data into actionable conservation insights.

2.3.2. Unstructured Data Cleaning

Filtering and improving the extensive vertebrate data from the Estonian Governmental Resources is the first step. This project specifically focuses on bird data, even though the original dataset encompasses all vertebrate species. In order to separate bird species from the larger dataset, filtering is done. This entails filtering by the 'Ruhm' column in EELIS data usage to make sure that only birds ('Linnud') are extracted.

The varied collection of PDFs obtained through focused web searches [3-5, 11-60] is covered in the next step. These documents frequently have inconsistent layouts and extraneous details that have no bearing on the conservation of birds. As a result, the text that is taken out of these PDFs is thoroughly cleaned. This entails removing non-informative sections, removing header and footer artifacts, and reformatting text to get rid of styles or graphical elements that aren't suitable for analysis. Only factual information that is directly related to bird conservation is kept thanks to this cleaning.

Lastly, the cleaned text data is carefully examined to guarantee consistency and clarity before text mining and the use of large language models. This entails aligning formatting, making sure that style and syntax are consistent, and looking for and standardizing species names to avoid discrepancies across documents. This cleaning process ensures that any unstructured data becomes well-suited for further usage.

2.3.3. LLM Usage

Text acquired via optical character recognition (OCR) undergoes a **post-OCR correction** process using the LLM. Post-OCR correction is the task of “correcting mistakes in optically identified texts from images or documents”. This step is essential for improving the quality and accuracy of the extracted text, as OCR systems often introduce errors such as misrecognized characters, broken words, or inconsistent formatting. Pretrained LLMs exhibit an inherent understanding of language that can be effectively leveraged to correct these mistakes [8].

Additionally, the LLM is essential for separating particular sections from each PDF document's table of contents. It finds and extracts pertinent sections related to the project, such as "Elupaik" (Habitat), "Elupaiga seisund" (State of Habitat), "Ohud" (Threats), "Populatsiooni muutused Eestis" (Population Changes in Estonia), "Kas rändlinnud" (Are They Migratory Birds), "Läbiviidud uuringud" (Conducted Studies), "Kavandatud uuringud" (Planned Studies), "Seisund ELis" (Status in the EU), and "Populatsiooni muutused teistes ELi riikides" (Population Changes in Other EU Countries), which later, combined, are used for the summarization task for the creation of insightful species general and threats-centered descriptions. This process ensures that only relevant ecological and conservation information is selected for in-depth analysis.

To ensure effective and consistent results, the prompt used for guiding the LLM was designed following best practices outlined by Zhang et al. (2024), where the model is instructed to take on a specific expert role—in this case, an *ornithology data assistant* tasked with extracting ornithological information. The prompt directly defined the role, task, and structure of the expected response, employing zero-shot strategies and structured context to enhance relevance and reasoning. As noted by Zhang et al., such prompt engineering techniques, including the use of role-based instruction and clear answer formatting, significantly improve the performance and interpretability of LLM outputs [9].

Finally LLM contributes to data extraction and summarization from these identified sections. For each category, including habitats, threats, and population dynamics, the LLM extracts essential information and synthesizes it into concise summaries. This process generates structured outputs that enhance clarity and provide insights into specific aspects of bird conservation, offering vital information for planning and decision-making.

3. RESULTS

3.1. Criteria And Methods For Technology Reliability Evaluation

The findings of this project are compiled in a CSV dataset, featuring the following columns: Estonian Name, Latin Name, Category, EELIS Link, Strategy Present, English Name, Group, Protection Category, Description, Description Rating (0–10), Threat Factor Description, Threat Factor Description Rating (0–10), Evaluator, Source Text Document Title, and Source Text Document.

The dataset is also accessible as a Google Sheets document and includes descriptions of 57 bird species categorized under various conservation statuses: I, II, or III. These descriptions are derived from bird conservation strategies and nature conservation documents that reference bird species in their research. Data entries lacking sufficient information from the available materials are denoted as 'NA'.

The project resulted in the creation of 40 bird descriptions and 38 threat descriptions that were not previously available in the EELIS system.

To assess the quality of the project's outcomes, an expert from the Eesti Ornitoloogiaühing was consulted. The expert evaluated the Description and Threat Factor Description based on quality and reliability, using a rating scale from 0 to 10 - where 0 signifies complete inaccuracy and 10 represents full accuracy. The expert provided ratings for five bird species within their area of expertise, among which two most informative were chosen. The second expert, evaluating the results in connection to the algorithm quality assessment and not general value in ornithology, gave a score of 1-10 on similar criteria for the two example articles given, but not in terms of how accurate the final information is in relation to the bird species in general, but only in relation to the original material (article or report).

For transparency, the evaluators was given the choice to either disclose their name or submit their evaluation anonymously; since evaluators chose the last option, their names are not mentioned. Sample results and their evaluations are presented in the subsequent section. Detailed output, with the results for all of the processed species, is available in Appendix 1.

3.2. Test Results

Four species of birds were selected for evaluation of the results:

- Väikeluik (Bewick's Swan):
- Rüüt (European Golden Plover):
- Laanerähn (Three-toed woodpecker)

- Hänilane (Western Yellow Wagtail)

Of these, for two species, the quality of the algorithm's performance on the basis of the original data was assessed, and for two species, the quality of the output data in general was evaluated. This allows the results to be evaluated from two perspectives at once - as efficiency in relation to the implemented solution (how well the created algorithm copes with data extraction and summarisation), and as efficiency in general as finished data (how relevant and useful the output data are).

3.2.1. Assessment of the quality of the algorithm's performance

To evaluate the reliability of the algorithm in extracting and summarizing information from source texts, two bird species were selected for detailed assessment: Rüüt (European Golden Plover) and Väikeluik (Bewick's Swan). The assessment focused on how accurately and clearly the algorithm summarized information from the original reports, without adding misleading details or omitting key points. The evaluation was based on a scale of 0 to 10, where a higher score indicates better performance.

For Rüüt, generated descriptions for bird information and threats descriptions were analyzed. The first summary (bird general information) received a score of 7/10, as it was mostly clear and well-structured but contained minor numerical inaccuracies and some generalized statements. One inaccuracy was the estimated number of breeding pairs, which was stated as 72 pairs, whereas the original text gave a range of 80–110 pairs (misrepresentation of key population data) [5]. Additionally, the summary included the phrase "world-famous riverine and open areas," which was not mentioned in the original and introduced unnecessary generalization (added imprecise or misleading details). The second, shorter summary (bird treats) was rated 6/10, as it omitted key details about the specific conservation measures and threats mentioned in the original text (lack of specificity). While the summary maintained the core ideas, the reduced detail level made it less useful for conservation planning.

For Väikeluik, one general description and one focusing on threats also were generated. The general description received a score of 7/10, as it successfully captured the overall migration patterns, habitat preferences, and conservation concerns. However, it omitted quantitative migration data, such as the exact number of stopover sites over time (e.g., "41 stopover sites in spring, 28 in autumn in the 1990s, reduced to 30 by 2017"), summarizing this instead as a "significant decrease in stopover sites" (loss of precise numerical data) [55]. Additionally, the description did not specify how feeding area changes led to a shift from western to central and eastern Estonia, instead stating that "feeding areas have changed", making it less useful.

The threats summary was rated 7/10, as it outlined key risk factors such as habitat destruction, disturbances, and collisions with man-made structures but lacked specific threat levels and priority rankings. The original text categorized threats into critical, high, moderate, and low impact, but this structured ranking was omitted in the summary, reducing the usefulness of the information (loss of structured risk assessment). Additionally, while the summary correctly mentioned collisions with power lines as a major threat, it did not specify the most dangerous locations, such as Väike Väin, where 48 dead swans were recorded within a single year (omission of location-based risk factors). The summary also generalized the effects of agricultural intensification, failing to highlight how changes in farmland use in different regions influenced food availability (loss of conservation-relevant land use information).

A key takeaway from the evaluation is that while the initial data extraction step (gathering relevant species information from PDFs) was effective and captured all necessary details, the

summarization step needs refinement. Specifically, the GPT-based summarization query should be improved with direct formulation from an ornithologist who can highlight the exact details required for conservation purposes. This expert-driven refinement would ensure that summaries align more closely with conservation standards and avoid the omission of essential ecological parameters.

Furthermore, the algorithm demonstrated equal effectiveness in handling different types of source materials. For Rüüt, the original text was extracted from a report covering multiple bird species in brief, whereas for Väikeluik, the original material was a comprehensive multi-page report dedicated entirely to the species. Despite these differences in document structure, the algorithm processed and summarized both types of sources with similar levels of accuracy and completeness. This suggests that the system can handle concise multi-species reports and extensive species-specific documents.

However, refining the summarization logic with expert oversight will be essential to improve the accuracy and specificity of the final outputs.

3.2.2. Assessment of the quality of the output data

To evaluate the quality of the algorithm's output data, two bird species were selected for assessment: Laanerähn (Three-toed Woodpecker) and Hänilane (Western Yellow Wagtail). The evaluation focused on the overall usefulness and completeness of the final summaries as conservation data, rather than their direct alignment with original source texts. The assessment considered how well the summaries captured key species information, conservation status, and threat factors, ensuring they provided accurate and structured insights for conservation planning. The evaluation was conducted on a scale of 0 to 10, where a higher score indicates greater accuracy, completeness, and relevance.

For Laanerähn (Three-toed Woodpecker), while the general summary correctly stated that Laanerähn is a resident species that depends on old-growth forests with sufficient deadwood, it failed to provide data on recent population changes or monitoring results, leading to a score of 5/10 for the general description. The description mentioned that most nesting areas are in strictly protected zones, but it did not clarify how effectively these zones are maintaining population stability [14]. Additionally, while habitat loss was identified as the main threat, the threat summary also received a 5/10, as it did not elaborate on the severity or impact of forestry practices, nor did it suggest concrete conservation measures beyond general habitat protection. The evaluator also noted that including regional conservation efforts, specific population trends, and structured risk assessments would significantly improve the summary's usefulness.

For Hänilane (Western Yellow Wagtail), the summary was too vague and lacked essential conservation details, resulting in a low evaluation score of 3/10 for both the general description and threat summary. While it mentioned that the species had been part of monitoring programs in 2014 and 2020 [4], it failed to provide specific population estimates or trends, making it difficult to assess the species' current status. The description lacked insights into key threats—such as habitat degradation due to agricultural intensification or climate-related changes—which are important for understanding the challenges it faces. Additionally, the threat summary hasn't specified in detail of what risks affect the species or how conservation actions should be prioritized. The evaluator also pointed out irrelevant or unclear phrases, such as the mention of "Sevena päeva loomadele," which does not contribute to the conservation context.

The evaluation demonstrates that while GPT-based summarization can extract and organize conservation-related information, expert input from ornithologists is still necessary to ensure accuracy, completeness, and relevance. The algorithm effectively identifies key conservation aspects but does not always capture detailed population trends, structured threat assessments, or specific conservation actions. As seen in the assessments of Laanerähn and Hänilane, the summaries provided general information but lacked important quantitative data and precise conservation measures.

Despite these limitations, the approach offers a practical way to streamline data processing, particularly when recent scientific papers and conservation reports are used as input. By automating the extraction and structuring of key details, the system can reduce the time spent on manual data compilation. However, ornithologists' involvement remains essential to refine the summaries by adding contextual details, the latest findings, and species-specific risk assessments, ensuring the information is suitable for conservation planning.

Estonian Name	Category	Kaitsekategooria	Kirjeldus	Kirjelduse hinnang (0-10)	Ohutegurite kirjeldus	Ohutegurite kirjelduse hinnang (0-10)	Algteksti dokumendi pealkiri	Evaluator Comments
laanerähn	II	II kategooria	Eestis mandriosas aastaringselt esinev sage haudelind. Laanerähn elab vanades looduslikes metsades, kus on rohkesti surnud või surevaid puid. Nigula looduskaitse alal pesitseb vähemalt 10 paar laanerähni.	5	Liigi peamiseks ohuteguriks on sobivate elupaikade hävimine. Kaitsekorral dusperioodi eesmärk on tagada laanerähni pesitsuseduk us ning vajadusel teha seireid, et jälgida liikide populatsioon i seisundit ja elupaikade kvaliteeti.	5		Laanerähn ehk kolmvarvas-r ähn (Picoides tridactylus) on Eestis mandriosas aastaringselt esinev sage haudelind, kelle arvukus ja populatsiooni seisund on kaitse-eesmä rkide hulka kantud. Nigula looduskaitsea lal pesitseb vähemalt 10 paar laanerähni. Valdav osa laanerähni pesapaiku on tsoneeritud

								sihtkaitsevöö ndisse, kus tuleb arvestada Looduskaitse seaduses sätestatud piiranguid. Laanerähn elab vanades looduslikes metsades ja sellele sobivad elupaigad on need, kus on rohkesti surnud või surevaid puid. Liigi peamiseks ohuteguriks on sobivate elupaikade hävimine, mistõttu on oluline kaitsta ja säilitada nende looduslikke elupaiku. Kaitsekorrald usperioodi eesmärk on tagada

								laanerähni pesitsuseduk us ning vajadusel teha seireid, et jälgida liikide populatsiooni seisundit ja elupaikade kvaliteeti.

hänilane	III	III kategooria	Hänilane (Motacilla flava) on olnud seotud erinevate seireaastatega, sealhulgas 2014. ja 2020. aastal. Audru poldri looduskaitsealal on linnuliikide arvukuse seire oluline osa, kus seiratakse lisaks hanedele ja luikedele ka muid rändlinde, sealhulgas hänilasi. Kaitse-eesmärkide hulka kuulub ka	3	Kaitse-eeskirja kohaselt tuleb tähelepanu pöörata linnukaitse tõhususele, et seal pesitsevatel lindudel, sealhulgas hanilastel, oleks tagatud rahulik ja turvaline elupaik. Hooldustööd on vajalikud, et tagada elupaikade säilimine ja edendada ohustatud liikide, näiteks kuldhänilase pesitsemise	3	Audru poldri looduskaitseala kaitsekorralduskava	Hänilane (Motacilla flava) on olnud seotud erinevate seireaastatega, sealhulgas 2014. ja 2020. aastal, kuid 2014. ja 2020. aastal ei ole selle liigi täpset arvu teada. Audru poldri looduskaitsealal on linnuliikide arvukuse seire olulise osa, kus seiratakse lisaks hanedele ja luikedele ka muid rändlinde, sealhulgas hänilasi. Kaitse-eesmärkide hulka kuulub ka selle liigi elupaikade

			selle liigi elupaikade kaitse. Pesitsejatel peavad olema sobivad tingimused pesitsemiseks, sealhulgas nutikalt planeeritud niidukooslus te hooldamine, karjatamine ja niitmine.		edendamist.			kaitse. Audru poldri alal peavad pesitsejatel olema sobivad tingimused pesitsemiseks, sealhulgas nutikalt planeeritud niidukooslust e hooldamine, karjatamine ja niitmine. Hooldustööd on vajalikud, et tagada elupaikade säilimine ja edendada ohustatud liikide, näiteks kuldhänilase pesitsemise edendamist. Kaitse-eeskirj a kohaselt tuleb ka tähelepanu pöörata linnukaitse

								tõhususele, et seal pesitsevatel lindudel, sealhulgas hanilastel, oleks tagatud rahulik ja turvaline elupaik. Sevena päeva loomadele tuleb tagada piisavalt eluruumi ja toitu, et toetada nende edasist elu ja rände ajaligoliksust. Kokkuvõtvalt on hänilane Audru poldri looduskaitseala l oluline liik, mille kaitse ja hoolsus seondub laiemate loodushoiu eesmärkideg a ning

								elupaikade hooldamise ja taastamise praktikatega.

Figure 6. *Results Table with Expert Ornithologist Evaluation*

CONCLUSION

The study successfully created and implemented a complete workflow, encompassing data collection from multiple Estonian sources, automated text extraction, LLM-driven summarization, and expert validation. This resulted in the in achieving all project objectives, and the generation of new structured descriptions for species and threats (**40 detailed bird descriptions and 38 threat descriptions**, which were previously unavailable in the EELIS system), filling critical gaps in existing databases. The system significantly reduces the time required for manual data compilation, providing a scalable approach to maintaining up-to-date ornithological records.

Expert evaluations highlighted both the strengths and areas for improvement of the system. The automated pipeline was effective in structuring large volumes of unprocessed conservation texts into coherent species descriptions, making it a valuable tool for ornithologists. However, the assessments also identified limitations, particularly in maintaining numerical accuracy, preserving specific threat classifications, and ensuring domain-specific linguistic precision. While the LLM-based summarization provided reliable general overviews, expert oversight remains necessary to refine outputs, particularly for conservation planning.

A key achievement of this work was the extensive engagement with the ornithological community, ensuring that real-world expertise informed both system design and output validation. Contacting and securing feedback from field specialists was essential in verifying the system's accuracy and relevance, reinforcing its practical contribution to Estonian biodiversity research.

Overall, this research demonstrates that LLMs can be successfully applied to ornithological data processing in a low-resource language. The system provides a great foundation for further refinement, with potential enhancements including improved summarization accuracy, deeper integration of structured threat assessments, and closer alignment with expert-driven conservation methodologies.

By covering such data gaps, the system contributes to more efficient and informed bird conservation efforts in Estonia.

RESUME

Uuringus loodi ja rakendati edukalt terviklik töövoog, mis hõlmas andmete kogumist mitmetest Eesti allikatest, automatiseeritud tekstiekstraktsiooni, LLM-põhist kokkuvõtete koostamist ja eksperthindamist. Selle tulemusena on kõik projekti seatud eesmärgid saavutatud ning loodud uued struktureeritud liikide ja ohtude kirjeldused (40 üksikasjalikku linnukirjeldust ja 38 ohukirjeldust), mis varem puudusid EELIS süsteemis), täites olulisi lünki olemasolevates andmebaasides. Süsteem vähendab märkimisväärselt käsitsi andmekoostamisele kuluvat aega, pakkudes skaleeritavat lähenemist ajakohaste ornitoloogiliste andmete haldamiseks.

Eksperthinnangud tõid esile nii süsteemi tugevused kui ka parendamist vajavad aspektid. Automatiseeritud töövoog osutus tõhusaks suurte töötlemata looduskaitsetekstide struktureerimisel sidusateks liikide kirjeldusteks, muutes selle väärtuslikuks tööriistaks ornitoloogidele. Samas toodi hinnangutes välja ka piirangud, eriti arvulise täpsuse säilitamisel, konkreetsete ohuklassifikatsioonide säilitamisel ja valdkonnaspetsiifilise keelelise täpsuse tagamisel. Kuigi LLM-põhised kokkuvõtted pakkusid usaldusväärseid üldiseid ülevaateid, on ekspertide järelvalve jätkuvalt vajalik tulemuste täpsustamiseks, eriti looduskaitse kavandamisel.

Üheks selle töö olulisemaks saavutuseks oli ulatuslik koostöö ornitoloogiakogukonnaga, tagades, et tegelik eksperditeave mõjutas nii süsteemi kujundust kui ka tulemuste valideerimist. Väliuuringute spetsialistide kaasamine ja nende tagasiside hankimine oli hädavajalik süsteemi täpsuse ja asjakohasuse kinnitamiseks, tugevdades selle praktilist panust Eesti elurikkuse uurimisse.

Kokkuvõttes näitab see uuring, et LLM-e saab edukalt rakendada ornitoloogiliste andmete töötlemisel väikese ressursiga keeles. Süsteem pakub tugeva aluse edasiseks arendamiseks, võimalikud täiustused hõlmavad kokkuvõtete täpsuse suurendamist, struktureeritud ohuhinnangute sügavamat integreerimist ja tihedamat kooskõla ekspertpõhiste looduskaitsemetoodikatega.

Selliste andmelünkade katmise kaudu aitab süsteem kaasa tõhusamatele ja teadlikumatele linnukaitsetegevustele Eestis.

REFERENCES

[1] EELIS (Eesti looduse infosüsteemi) infoleht, https://infoleht.keskkonnainfo.ee/ (05.08.2024)

[2] Botaurus stellaris taksonikirjeldus, https://app.plutof.ut.ee/taxon-description/view/132 (01.04.2017)

[3] Luitemaa looduskaitseala ja Luitemaa hoiuala kaitsekorralduskava 2018-2027, https://keskkonnaamet.ee/sites/default/files/documents/2021-06/Luitemaa%20kaitsekorralduskava%2012.04.2021%20redaktsioon.pdf (12.04.2021)

[4] Audru poldri looduskaitseala kaitsekorralduskava, https://keskkonnaamet.ee/sites/default/files/documents/2023-04/Audru%20poldri%20looduskaitseala%20kaitsekorralduskava%20%28muudetud%29_0.pdf (30.03.2023)

[5] Alam-Pedja linnu- ja loodusala kaitsekorralduskava 2016-2025, https://rsis.ramsar.org/RISapp/files/49164532/documents/EE905_mgt200114.pdf (29.08.2017)

[6] Lochan Basyal, Mihir Sanghvi (2023). Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. https://arxiv.org/abs/2310.10449 (Last accessed: 17.10.2023)

[7] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch (2024). From Text to Insight: Large Language Models for Materials Science Data Extraction. https://www.researchgate.net/publication/382527116_From_Text_to_Insight_Large_Language_Models_for_Materials_Science_Data_Extraction (Last accessed: 01.07.2024)

[8] Martijn Veninga (2024). LLMs for OCR Post-Correction https://essay.utwente.nl/102117/1/Veninga_MA_EEMCS.pdf (Last accessed: 01.07.2024)

[9] Haochen Zhang, Yuyang Dong, Chuan Xiao, Masafumi Oyamada (2024). Large Language Models as Data Preprocessors. https://arxiv.org/abs/2308.16361 (Last accessed: 27.10.2024)

[10] Keskkonnaamet. Paljassaare hoiuala kaitsekorralduskava 2020–2029. https://keskkonnaamet.ee/sites/default/files/documents/2021-06/A2%20Paljassaare%20kaitsekorralduskava%2C%20projektiala%2038.pdf (Last accessed: 01.06.2021)

[11] Keskkonnaamet (Martijn van Schie, Margus Ellermaa). Audru poldri looduskaitseala kaitsekorralduskava. https://keskkonnaamet.ee/sites/default/files/documents/2023-04/Audru%20poldri%20looduskaitseala%20kaitsekorralduskava%20%28muudetud%29_0.pdf (Last accessed: 01.01.2023)

[12] Keskkonnaamet. Alam-Pedja linnu- ja loodusala kaitsekorralduskava 2016-2025. https://rsis.ramsar.org/RISapp/files/49164532/documents/EE905_mgt200114.pdf (Last accessed: 29.08.2017)

[13] Keskkonnaamet. Peipsiveere looduskaitseala kaitsekorralduskava 2016-2025. https://kaitsealad.ee/sites/default/files/inline-files/PEIPSIVEERE_LKA_KKK_2016_2025.pdf (Last accessed: 01.01.2016)

[14] Keskkonnaamet. Nigula looduskaitseala kaitsekorralduskava 2015-2024. https://infoleht.keskkonnainfo.ee/ala/1020 (Last accessed: 01.01.2015)

[15] Keskkonnaamet. Luitemaa looduskaitseala ja Luitemaa hoiuala kaitsekorralduskava 2018-2027. https://keskkonnaamet.ee/sites/default/files/documents/2021-06/Luitemaa%20kaitsekorralduskava%2012.04.2021%20redaktsioon.pdf (Last accessed: 01.01.2018)

[16] Keskkonnaamet. Lihula maastikukaitseala ja Lihula hoiuala kaitsekorralduskava 2016–2025. https://rsis.ramsar.org/RISapp/files/49174441/documents/EE1997_mgt200110.pdf?language=en (Last accessed: 01.01.2016)

[17] Keskkonnaamet. Silma looduskaitseala ja Karjatsimere hoiuala kaitsekorralduskava 2017–2026. https://rsis.ramsar.org/RISapp/files/53381440/documents/EE2022_mgt180119_1.pdf (Last accessed: 01.01.2017)

[18] Roasto, R., Tampere, U. Eesti looduse kaitse aastal 2020 https://loodusveeb.ee/sites/default/files/inline-files/elk_2020_est.pdf (Last accessed: 01.01.2020)

[19] Keskkonnaamet. Lihula maastikukaitseala ja Lihula hoiuala kaitsekorralduskava 2016–2025. https://rsis.ramsar.org/RISapp/files/49174441/documents/EE1997_mgt200110.pdf?language=en (Last accessed: 01.01.2016)

[20] Keskkonnaamet. Hanede ja laglede kaitse ja ohjamise tegevuskava. https://www.keskkonnaamet.ee/sites/default/files/hanede_tk_2021-2025.pdf (Last accessed: 22.03.2021)

[21] Keskkonnaamet. Kaisma loodusala kaitsekorralduskava. https://keskkonnaamet.ee/sites/default/files/documents/2023-12/Kaisma%20loodusala%20kaitsekorralduskava.pdf (Last accessed: 15.12.2023)

[22] Indrek Tammekänd, EOÜ. Kaitstavate linnuliikide pesitsuselupaikade kaardistamise ja Eelisesse vormistamise põhimõtted. https://eoy.ee/img/Kaitstavate_linnuliikide_pesitsuselupaikade_kaardistamise_ja_EELISesse_vormistamise_pohimotted.pdf (Last accessed: 22.02.2023)

[23] Keskkonnaamet. Kalakotka (Pandion haliaetus) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2021-05/kalakotka_ltk_28102019.pdf (Last accessed: 12.11.2019)

[24] Keskkonnaamet. Kaljukotka kaitse tegevuskava. https://www.kotkas.ee/files/kaljukotka_ktk_2018-2022.pdf (Last accessed: 03.12.2018)

[25] Ülo Väli, Aarne Tuule (Keskkonnaamet). Kanakulli (Accipiter gentilis) kaitse tegevuskava. https://eelis.ee/getdok/-1915000682 (Last accessed: 01.01.2018)

[26] Renno Nellis, Rein Nellise, Madis Leivitsa, Urmas Sellise, Aarne Tuule, Raivo Endreksoniga. (Keskkonnaamet). Kassikaku (Bubo bubo) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2024-04/Kassikaku%20kaitse%20tegevuskava.pdf (Last accessed: 18.04.2024)

[27] Keskkonnaamet. Tündre loodusala (Tündre looduskaitseala) kaitsekorralduskava. https://keskkonnaamet.ee/sites/default/files/documents/2024-04/KeA_LISA_kavade_kinnitamine%20%283%29.pdf (Last accessed: 03.04.2024)

[28] Ülo Väli. Suur-konnakotka kaitse tegevuskava aastateks 2006–2010. https://www.kotkas.ee/failid/KKK_AC.pdf (Last accessed: 01.01.2005)

[29] Ülo Väli. Laanepüü (Bonasa bonasia) kaitse tegevuskava 2015–2019 (2018)

[30] Eesti Ornitoloogiaühing, Kotkaklubi. Üle-eestiline maismaalinnustiku analüüs. https://kliimaministeerium.ee/sites/default/files/documents/2022-12/L%C3%B5pparuanne%20-%20%C3%9Cle-eestiline%20maismaalinnustiku%20anal%C3%BC%C3%BCs_0.pdf (Last accessed: 01.01.2022)

[31] Riho Marja, Jaanus Elts, Joosep Tuvi, James Phillips. Rukkirääga (Crex crex) arvukuse varieeruvus elupaigatüüpide lõikes Lahemaa rahvuspargis 2014. aastal. https://hirundo.eoy.ee/file_download/407/Marja_et_al_2015-2.pdf (Last accessed: 01.02.2015)

[32] Keskkonnaamet. Merikotka (Haliaeetus albicilla) kaitse tegevuskava. https://kliimaministeerium.ee/sites/default/files/documents/2022-12/L%C3%B5pparuanne%20-%20%C3%9Cle-eestiline%20maismaalinnustiku%20anal%C3%BC%C3%BCs_0.pdf (Last accessed: 01.01.2019)

[33] Keskkonnaamet. Metsise (Tetrao urogallus) kaitse tegevuskava. https://loodusveeb.ee/sites/default/files/inline-files/Metsis%202015.pdf (Last accessed: 01.01.2015)

[34] Keskkonnaamet. Muraka loodus- ja linnuala kaitsekorralduskava. https://keskkonnaamet.ee/sites/default/files/documents/2024-02/Muraka%20loodus-%20ja%20linnuala%20kkk%202024.pdf (Last accessed: 02.02.2024)

[35] Keskkonnaamet. Must-toonekure (Ciconia nigra) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2021-05/must_toonekure_kaitse_tegevuskava.pdf (Last accessed: 14.02.2018)

[36] Keskkonnaamet. Mustviire (Chlidonias niger) kaitse tegevuskava. (10.07.2015)

[37] Keskkonnaamet. Nabala-Tuhala looduskaitseala kaitsekorralduskava 2023–2032. https://keskkonnaamet.ee/sites/default/files/documents/2022-10/Nabala-Tuhala%20looduskaitseala%20kaitsekorralduskava%202023-2032.pdf (Last accessed: 05.10.2022)

[38] Keskkonnaamet. Niidurüdi (Calidris alpina schinzii) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2023-12/niidurydi_kaitse_tegevuskava.pdf (Last accessed: 28.12.2023)

[39] Keskkonnaamet. Pakri loodus- ja linnuala (Pakri hoiuala, Pakri maastikukaitseala) kaitsekorralduskava. https://eelis.ee/kava/2109737842 (Last accessed: 05.04.2024)

[40] Keskkonnaamet. Põduste luha loodusala kaitsekorralduskava. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://keskkonnaamet.ee/media/7327/download (Last accessed: 20.12.2023)

[41] Keskkonnaamet. Põldtsiitsitaja (Emberiza hortulana) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2021-05/poldtsiitsitaja_ltk.pdf (Last accessed: 07.05.2020)

[42] Keskkonnaamet. Puhatu loodus- ja linnuala kaitsekorralduskava. https://keskkonnaamet.ee/sites/default/files/documents/2024-04/Puhatu%20linnu-%20ja%20loodusala%20kaitsekorralduskava%202024.pdf (Last accessed: 05.04.2024)

[43] Keskkonnaamet. Punaselg-õgija (Lanius collurio) kaitse tegevuskava. (Last accessed: 10.07.2015)

[44] Keskkonnaamet. Rohunepi (Gallinago media) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2021-07/Rohunepi%20tegevuskava.pdf (Last accessed: 29.06.2021)

[45] Keskkonnaamet. Roostikulindude kaitse tegevuskava. (Last accessed: 14.04.2015)

[46] Keskkonnaamet. Siiraku loodusala kaitsekorralduskava. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://keskkonnaamet.ee/media/7433/download (Last accessed: 22.12.2023)

[47] Keskkonnaamet. Sookure (Grus grus) kaitse tegevuskava. (Last accessed: 24.07.2015)

[48] Keskkonnaamet. Soomaa rahvuspargi, Soomaa loodusala ja Soomaa linnuala kaitsekorralduskava 2022–2031. https://keskkonnaamet.ee/sites/default/files/documents/2022-03/Soomaa%20RP%20KKK%202022-2031.pdf (Last accessed: 23.09.2023)

[49] Keskkonnaamet. Suur-konnakotka (Clanga clanga) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2024-04/Suur-konnakotka%20tegevuskava.pdf (Last accessed: 04.04.2024)

[50] Keskkonnaamet. Tiirude kaitse tegevuskava. (Last accessed: 03.09.2015)

[51] Keskkonnaamet. Tudusoo loodus- ja linnuala kaitsekorralduskava. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://keskkonnaamet.ee/media/7616/download (Last accessed: 02.02.2024)

[52] Keskkonnaamet. Uhtju loodusala kaitsekorralduskava. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://keskkonnaamet.ee/media/6937/download (Last accessed: 27.09.2023)

[53] Keskkonnaamet. Väike-konnakotka (Aquila pomarina) kaitse tegevuskava. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://keskkonnaamet.ee/sites/default/files/documents/2021-05/vaike-konnakotka_tk.pdf (Last accessed: 26.03.2018)

[54] Keskkonnaamet. Väike-laukhane (Anser erythropus) kaitse tegevuskava. (Last accessed: 24.07.2015)

[55] Keskkonnaamet. Väikeluige (Cygnus columbianus bewickii Yarr.) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2021-05/vaikeluige_ktk_14.03.2018.pdf (Last accessed: 18.04.2018)

[56] Keskkonnaamet. Valgeselg-kirjurähni (Dendrocopos leucotos) kaitse tegevuskava. (Last accessed: 2012)

[57] Keskkonnaamet. Villtulika (Ranunculus lanuginosus L.) kaitse tegevuskava. https://keskkonnaamet.ee/sites/default/files/documents/2022-08/Villtulikas_LTK_avalik_1.pdf (Last accessed: 24.08.2022)

[58] Rein Nellis, Keskkonnaamet. Väike-kärbsenäpi (Ficedula parva) kaitse tegevuskava. (Last accessed: 10.09.2015)

[59] Keskkonnaamet. Laanepüü (Bonasa bonasia) kaitse tegevuskava. (Last accessed: 12.03.2015)

APPENDIX 1

Table with the detailed results:
https://docs.google.com/spreadsheets/d/15m0kZQUyaPa72ga5_ckwAq_fvWIV8_61ITF4-9NsQBI/edit?usp=sharing